DEPARTMENT OF STATISTICS

THE WHARTON SCHOOL

**University of Pennsylvania**

Statistics 405X/705X Quarter 4 Spring 2016

# Statistical computing with R

Syllabus

---

**PEOPLE:**

Instructor: Richard Waterman waterman@wharton.upenn.edu 443 JMHH

Classes meet: Tuesday/Thursday from 12-1:30PM in 245 JMHH.

Office hours: Tuesday/Thursday from 1:30-3:00PM.

Teaching Assistant: TBD

Course website: All course related materials will be distributed via the Canvas website and grades can be checked during the Quarter using Canvas. You can also use the discussion feature on Canvas to ask questions regarding the course material, assignments and any scheduling issues.

## BACKGROUND

The R statistical programming environment has long been the platform of choice for quantitative statistical research activities. Literally, thousands of add-on packages are available providing for a comprehensive suite of data analysis tools. Recently, R has made a major push into the business environment where it is frequently used by Data Scientists. For example, Oracle has now has an in-database version of R, "R-enterprise" taking advantage of parallelization and is suitable for "big data" type problems. The popular Tableau visualization software also has an interface for R.

The goal of this course is to introduce students to the R programming environment and related eco-system and thus provide them with an in-demand skill-set, in both the research and business environments. Further, R is a platform that is used in other advanced classes taught at Wharton, so that this class will prepare students for these higher level classes and electives.

One key feature of R is that it is Open Source and is freely distributed. Consequently, it is a platform, that once learnt, will remain universally available to students.

## COURSE OVERVIEW

This one quarter course will expose students to the R statistical programming language. No previous programming experience is assumed. There will be an assumption that students have completed prior statistics courses to the level of multiple regression analysis. Specific acceptable prerequisites are listed below.

Students will be expected to have reviewed the class notes prior to each class and will be expected to bring their laptops, loaded with R and the RStudio IDE to class. Classes will have a four part structure:

1. A topic overview
2. Instructor demonstration
3. Student hands-on group-based project activity
4. Wrap-up and project review

By the end of the course students should be able to code statistical functions in R. They should be able to extend the functionality of R by using add-on packages and they should be able to use R to perform the work-horse statistical tasks such as multiple-regression and simulation analyses.

## PRE-REQUISITES

Any of Stat 102, Stat 112, Stat 431, Stat 613, Stat 621 (or waiving the MBA statistics course). These classes all take students to the level of multiple regression.

## COURSE MATERIALS

*SOFTWARE:*

1. The R statistical software program. Available from: https://www.r-project.org/
2. RStudio an Integrated Development Environment (IDE) for R. Available from: https://www.rstudio.com/

*CLASS NOTES:* these will be available from Canvas.

There is no required textbook, though there are many optional texts available for students to refer to. Two recommended ones are: Introductory Statistics with R, Peter Dalgaard and An Introduction to Statistical Learning: with Applications in R, Gareth James et. al.

There are many quality free on-line tutorials and resources for R such as DataCamp: https://www.datacamp.com/courses/free-introduction-to-r that students may find useful.

## HOMEWORK

There will be 5 homeworks during the quarter. These homeworks will be prescriptive and involve performing a set of programming related tasks in R. The deliverables will be R code, output and related discussions. There is no final exam but rather a take home final project.

Homework should be submitted to Canvas as text files for code and PDF files for output.

## CLASS SCHEDULE

The class content is structured along the lines of using R for a project in a business environment. Specifically, there needs to be a problem definition and scoping stage. Data needs to be identified and read into the analysis platform. The analysis occurs. Results are reported to interested parties.

**Table 1 Class schedule**

| | |
|---|---|
| Module 1 | Introduction to R and RStudio. Using the help facility. |
| Module 2 | Data structures: vectors, matrices, lists and data frames. |
| Module 3 | Reading data into R from various data sources. Merging data across data sources. |
| Module 4 | Statistical modeling functions: lm and glm. |
| Module 5 | Writing your own functions I. |
| Module 6 | Writing your own functions II. |
| Module 7 | Iterating with R: logic and flow control. |
| Module 8 | Simulation I. |
| Module 9 | Simulation II. |
| Module 10 | Extending R with add-on packages and the R ecosystem. |
| Module 11 | Graphics. |
| Module 12 | Dynamic and web reporting: Knitr and Shiny. Running R as part of a business pipeline—the R terminal. |

## CLASS CONTENT

### *MODULE 1.* Introduction to R and RStudio

In this class we will get to know R. This involves first of all installing R and RStudio.

The basic functionality of R will be demonstrated. Using R for calculations. Using R to calculate summary statistics on data. Using R to generate random numbers. Variable types in R. Numeric variables, strings and factors. Accessing the help system.

### *MODULE 2.* Data structures: vectors, matrices, lists and data frames

R makes extensive use of various data structures. The core data structures are vectors, matrices, arrays, lists and dataframes. We will discuss accessing elements of these data structures, sub-setting vectors, slicing arrays and drilling down on lists. We will also take a look at the apply and lapply functions, that allow you to apply functions to arrays and lists.

### *MODULE 3.* Reading data into R from various data sources. Merging data across data sources

R has many options for bringing in data for analysis. These include reading from flat files, reading from database connections and reading from web sources. Many problems involve multiple data sources, so we will discuss merging data sources in R using the join command.

### *MODULE 4.* **Statistical modeling functions: lm and glm**

Linear and generalized linear models (for example, logistic regression) are the workhorses of modern analytics. This class will illustrate the implementation of these functions in R and requires the use of the formula syntax for model specification. We will discuss prediction and model checking via residuals.

### *MODULE 5.* **Writing your own functions (I)**

One of the most powerful features of R is the ability to write your own functions. These functions may help pre-process data or implement specific computational algorithms. In this class we will introduce the R function syntax and in particular the passing of variables into the function, and argument handling.

### *MODULE 6.* **Writing your own functions (II)**

There are always elegant ways to write functions and more brute force approaches. We will describe approaches that make R more efficient in function evaluations, in particular vectorization. We will discuss the "…" notation that allows arguments to be passed on to other functions. We will discuss functions that themselves take other functions as arguments.

### *MODULE 7.* **Iterating with R. Logic and flow control**

In order to prepare for simulation modeling it is important to be able to control the flow of an algorithm. These functions include the if, for, while and break constructs.

### *MODULE 8.* **Simulation I**

This class will introduce Bootstrapping and Monte-Carlo simulation in R. We will investigate the sample command and be introduced to random number generation from the classic statistical probability distribution functions.

### *MODULE 9.* **Simulation II**

Expanding on the simulation ideas from the previous class we will investigate some permutation tests as alternatives to classical hypothesis tests .We will also use simulation to check whether or not modeling assumptions appear reasonable, for example whether normality for a quantity such as the maximum likelihood estimator is reasonable. Finally we will compare the efficiency of different sampling methodologies, simple random

sampling and stratified random sampling in terms of the efficiency of the estimates they produce.

### *MODULE 10.* Extending R with add-on packages and the R ecosystem

One of the great benefits of using R is the access to hundreds of add-ion packages. This class will discuss the R ecosystem, and illustrate R's extensibility. We will illustrate this extensibility through an example using the randomForest package and various other packages that work in conjunction with it.

### *MODULE 11.* Graphics

A picture is worth a thousand words and any statistical analysis with no graphics tends to fall a little flat. R has a wide range of graphic abilities, from the low level graphical primitives that can be used to build more complicated graphics to high level routines such as lattice and ggplot2. This class will explore some of these graphical capabilities.

### *MODULE 12.* Dynamic and web reporting: Knitr and Shiny. Running R as part of a business pipeline -- the R terminal

The sincerest form of flattery is implementation. In this class we will look at the facilities available to create living reports and on-line presentations driven by the R language. We will also discuss the Rterm shell that allows R to be run as a batch process and is useful in embedding R into a business work flow.

## GRADING

The final grade will be weighted using 75% from the five assignment (each counts as 15%) and 25% from the final project. All assignments will be included in the final grade. There is **no** "drop the lowest score" policy. Grade queries must be submitted within one week of the solutions being posted.

## CLASSROOM EXPECTATIONS

There is no formal participation component to the final grade but questions are strongly encouraged. Phones, laptops and other electronic devices are not to be used in class except during classroom project activities.